Chapter 8

Correlation

Concentrations of atrazine and nitrate in shallow groundwaters are measured in wells over a several county area. For each sample, the concentration of one is plotted versus the concentration of the other. As atrazine concentrations increase, so do nitrate. How might the strength of this association be measured and summarized?

Streams draining the Sierra Nevada mountains in California usually receive less precipitation in November than in other months. Has the amount of November precipitation significantly changed over the last 70 years, showing a gradual change in the climate of the area? How might this be tested?

The above situations require a measure of the strength of association between two continuous variables, such as between two chemical concentrations, or between amount of precipitation and time. How do they co-vary? One class of measures are called correlation coefficients, three of which are discussed in this chapter. Also discussed is how the significance of that association can be tested for, to determine whether the observed pattern differs from what is expected due entirely to chance. For measurements of correlation between grouped (non-continuous) variables, see Chapter 14.

Whenever a correlation coefficient is calculated, the data should be plotted on a scatterplot. No single numerical measure can substitute for the visual insight gained from a plot. Many different patterns can produce the same correlation coefficient, and similar strengths of relationships can produce differing coefficients, depending on the curvature of the relationship. In Chapter 2, figure 2.1 presented eight plots all with a linear correlation coefficient of 0.70. Yet the data were radically different! Never compute correlation coefficients and assume the data look like those in h of figure 2.1.

8.1 Characteristics of Correlation Coefficients

Correlation coefficients measure of the strength of association between two continuous variables. Of interest is whether one variable generally increases as the second increases, whether it decreases as the second increases, or whether their patterns of variation are totally unrelated. Correlation measures observed co-variation. It does not provide evidence for causal relationship between the two variables. One may cause the other, as precipitation causes runoff. They may also be correlated because both share the same cause, such as two solutes measured at a variety of times or a variety of locations. (Both are caused by variations in the source of the water). Evidence for causation must come from outside the statistical analysis -- from the knowledge of the processes involved.

Measures of correlation (here designated in general as ρ) have the characteristic of being dimensionless and scaled to lie in the range $-1 \leq \rho \leq 1$. When there is no correlation between two variables, $\rho = 0$. When one variable increases as the second increases, ρ is positive. When they vary in opposite directions, ρ is negative. The significance of the correlation is evaluated using a hypothesis test:

H_0: $\rho = 0$ versus H_1: $\rho \neq 0$.

When one variable is a measure of time or location, correlation becomes a test for temporal or spatial trend.

8.1.1 Monotonic Versus Linear Correlation

Data may be correlated in either a linear or nonlinear fashion. When y generally increases or decreases as x increases, the two variables are said to possess a monotonic correlation. This correlation may be nonlinear, with exponential patterns, piecewise linear patterns, or patterns similar to power functions when both variables are non-negative. Figure 8.1 illustrates a non-linear monotonic association between two variables -- as x increases, y generally increases by an ever-increasing rate. This nonlinearity is evidence that a measure of linear correlation would be inappropriate. The strength of a linear measure will be diluted by nonlinearity, resulting in a lower correlation coefficient and less significance than a linear relationship having the same amount of scatter.

Three measures of correlation are in common use -- Kendall's tau, Spearman's rho, and Pearson's r. The first two are based on ranks, and measure all monotonic relationships such as that in figure 8.1. They are also resistant to effects of outliers. The more commonly-used Pearson's r is a measure of linear correlation (figure 8.2), one specific type of monotonic correlation. None of the measures will detect nonmonotonic relationships, where the pattern doubles back on itself, like that in figure 8.3.

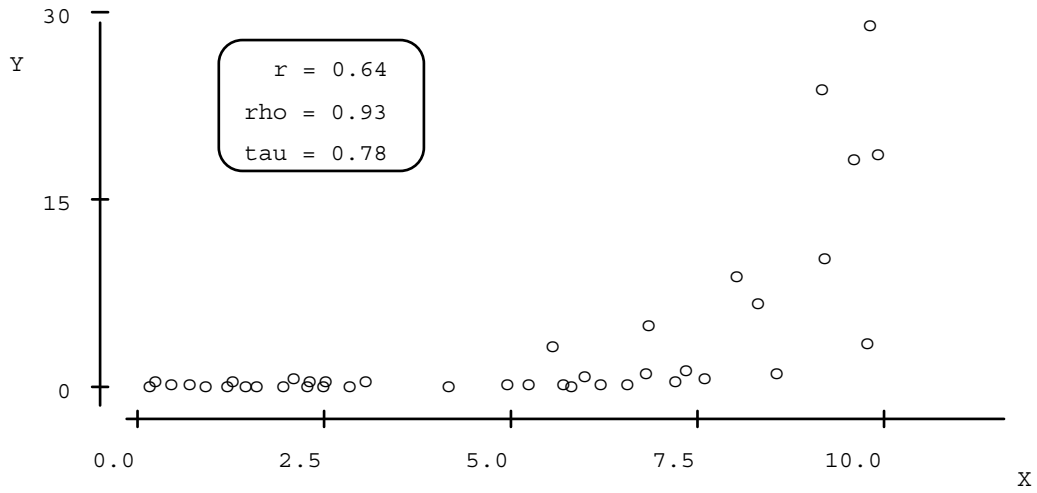

Figure 8.1 Monotonic (nonlinear) correlation between x and y.

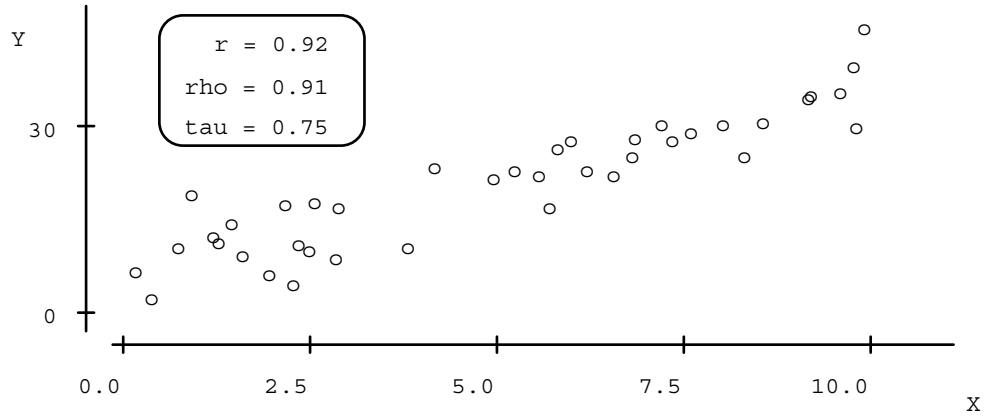

Figure 8.2 Linear correlation between X and Y.

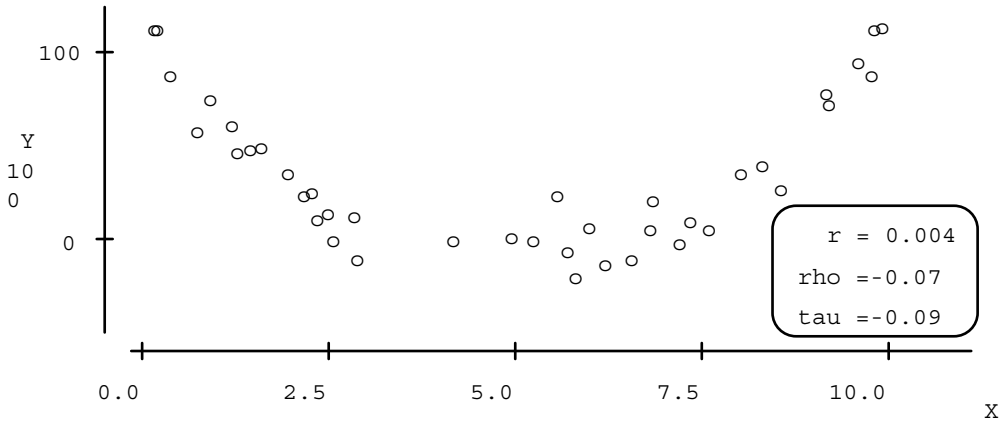

Figure 8.3 Non-monotonic relationship between X and Y.

8.2 Kendall's Tau

Tau (Kendall, 1938 and Kendall, 1975) measures the strength of the monotonic relationship between x and y. Tau is a rank-based procedure and is therefore resistant to the effect of a small number of unusual values. It is well-suited for variables which exhibit skewness around the general relationship.

Because tau (τ) depends only on the ranks of the data and not the values themselves, it can be implemented even in cases where some of the data are censored, such as concentrations known only as less than the reporting limit. This is an important feature of the test for applications to water resources. See Chapter 13 for more detail on analysis of censored data.

Tau will generally be lower than values of the traditional correlation coefficient r for linear associations of the same strength (figure 8.2). "Strong" linear correlations of 0.9 or above correspond to tau values of about 0.7 or above. These lower values do not mean that tau is less sensitive than r, but simply that a different scale of correlation is being used. Tau is easy to compute by hand, resistant to outliers, and measures all monotonic correlations (linear and nonlinear). Its large sample approximation produces p-values very near exact values, even for small sample sizes. As it is a rank correlation method, tau is invariant to monotonic power transformations of one or both variables. For example, τ for the correlation of log(y) versus log(x) will be identical to that of y versus log(x), and of y versus x.

8.2.1 Computation

Tau is most easily computed by first ordering all data pairs by increasing x. If a positive correlation exists, the y's will increase more often than decrease as x increases. For a negative correlation, the y's will decrease more often than increase. If no correlation exists, the y's will increase and decrease about the same number of times.

A two-sided test for correlation will evaluate the following equivalent statements for the null hypothesis H_0, as compared to the alternate hypothesis H_1:

H_0:
- a) no correlation exists between x and y ($\tau = 0$), or
- b) x and y are independent, or
- c) the distribution of y does not depend on x, or
- d) Prob ($y_i < y_j$ for $i < j$) = 1/2.

H_1:
- a) x and y are correlated ($\tau \neq 0$), or
- b) x and y are dependent, or
- c) the distribution of y (percentiles, etc.) depends on x, or
- d) Prob ($y_i < y_j$ for $i < j$) $\neq$ 1/2.

The test statistic S measures the monotonic dependence of y on x. Kendall's S is calculated by subtracting the number of "discordant pairs" M, the number of (x,y) pairs where y decreases as x increases, from the number of "concordant pairs" P, the number of (x,y) pairs where y increases with increasing x:

$$S = P - M \qquad [8.1]$$

where P = "number of pluses", the number of times the y's increase as the x's increase, or the number of $y_i < y_j$ for all $i < j$,

M = "number of minuses," the number of times the y's decrease as the x's increase, or the number of $y_i > y_j$ for $i < j$.

for all i = 1,....(n − 1) and j = (i+1),.....n.

Note that there are n(n−1)/2 possible comparisons to be made among the n data pairs. If all y values increased along with the x values, S = n(n−1)/2. In this situation, the correlation coefficient τ should equal +1. When all y values decrease with increasing x, S = −n(n−1)/2 and τ should equal −1. Therefore dividing S by n(n−1)/2 will give a value always falling between −1 and +1. This then is the definition of τ, measuring the strength of the monotonic association between two variables:

Kendall's tau correlation coefficient

$$\tau = \frac{S}{n(n-1)/2} \qquad [8.2]$$

To test for significance of τ, S is compared to what would be expected when the null hypothesis is true. If it is further from 0 than expected, H_0 is rejected. For $n \leq 10$ an exact test should be computed. The table of exact critical values is found in table B8 of the Appendix.

8.2.2 Large Sample Approximation

For n > 10 the test statistic can be modified to be closely approximated by a normal distribution. This large sample approximation Z_S is the same form of approximation as used in Chapter 5 for the rank-sum test, where now

d = 2 (S can vary only in jumps of 2),

μ_S = 0, and

$\sigma_S = \sqrt{(n/18)\bullet(n-1)\bullet(2n+5)}$.

$$Z_S = \begin{cases} \dfrac{S-1}{\sigma_s} & if\ S>0 \\ 0 & if\ S=0 \\ \dfrac{S+1}{\sigma_s} & if\ S<0 \end{cases} \qquad [8.3]$$

The null hypothesis is rejected at significance level α if $|Z_S| > Z_{crit}$ where Z_{crit} is the value of the standard normal distribution with a probability of exceedance of $\alpha/2$. In the case where some of the x and/or y values are tied the formula for σ_S must be modified, as discussed in the next section.

Example 1: 10 pairs of x and y are given below, ordered by increasing x:

y	1.22	2.20	4.80	1.28	1.97	1.46	2.64	2.34	4.84	2.96
x	2	24	99	197	377	544	632	3452	6587	53,170

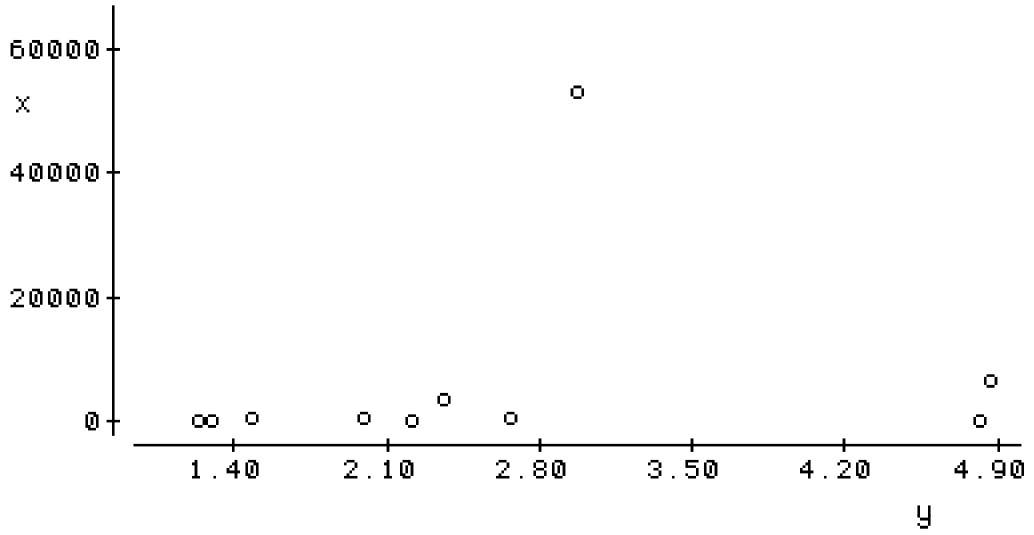

Figure 8.4 Example 1 data showing one outlier present.

To compute S, first compare y_1= 1.22 with all subsequent y's (y_j , j>1).

2.20 > 1.22, so score a +
4.80 > 1.22, score a +
1.28 > 1.22, score a +
1.97 > 1.22, score a + etc.

All subsequent y's are larger, so there are 9 +'s for i=1.

Move on to i=2, and compare y_2 =2.20 to all subsequent y's.

4.80 > 2.20, so score a +
1.28 < 2.20, score a −
1.97 < 2.20, score a −
1.46 < 2.20, score a − etc.

There are 5 +'s and 3 −'s for i=2. Continue in this way, until the final comparison of y_{n-1} = 4.84 to y_n. It is convenient to write all +'s and −'s below their respective y_i, as below:

y_i	1.22	2.20	4.80	1.28	1.97	1.46	2.64	2.34	4.84	2.96
	+	+	−	+	−	+	−	+	−	
	+	−	−	+	+	+	+	+		
	+	−	−	+	+	+	+			
	+	−	−	+	+	+				
	+	+	−	+	+					
	+	+	+	+						
	+	+	−							
	+	+								
	+									

In total there are 33 +'s (P = 33) and 12 −'s (M = 12). Therefore S = 33 − 12 = 21.
There are 10•9/2 = 45 possible comparisons, so $\tau = 21/45 = 0.47$.
Turning to table B8, for n=10 and S=21, the exact p-value is 2•0.036 = 0.072.

The large sample approximation is

$$Z_S = (21-1) / \sqrt{(10/18)\bullet(10-1)\bullet(20+5)}$$
$$= 20/(11.18) = 1.79.$$

From a table of the normal distribution, the 1-sided quantile for 1.79 = 0.963
so that $p \cong 2 \bullet (1-.963) = 0.074$

8.2.3 Correction for Ties

To compute τ when ties are present, tied values of **either x or y** produce a 0 rather than + or − . Ties do not contribute to either P or M. S and τ are computed exactly as before. An adjustment is required for the large sample approximation Z_S , however, by correcting the σ_S formula.

In order to compute σ_S in the presence of ties, both the number of ties and the number of values involved in each tie must be counted. Consider for example a water quality data set (in units of μg/L) of 17 values (n=17) shown here in ascending order.

<1, <1, <1, <1, <1, 2, 2, 2, 3, 5, 5, 7, 9, 10, 10, 14, 18.

There are a total of 4 tied groups in the data set. The largest tied group in the data set is of 5 values (tied at <1 μg/L), there are no tied groups of 4, there is 1 tied group of 3 (at 2 μg/L), and there are 2 tied groups of 2 (at 5 and 10 μg/L). For completeness note that there are 5 "ties" of extent 1 (untied values at 3, 7, 9, 14, and 18 μg/L). These appropriately never add to the

correction because (i–1) always equals zero. Kendall (1975) defined the variable t_i as the number of ties of extent i. For this data set $t_5 = 1$ (1 tie of extent 5), $t_4 = 0$ (no ties of extent 4), $t_3 = 1$ (1 tie of extent 3), $t_2 = 2$ (2 ties of extent 2) and $t_1 = 5$ (5 "ties" of extent 1). For i>5, $t_i = 0$. Kendall's correction to σ_S in the presence of ties is:

$$\sigma_S = \sqrt{\frac{[n\,(n-1)\,(2n+5) - \sum_{i=1}^{n} t_i\,(i)\,(i-1)\,(2i+5)\,]}{18}} \qquad [8.4]$$

So for the example water quality data:

$$\sigma_S = \sqrt{[17 \bullet 16 \bullet 39 - 5 \bullet 1 \bullet 0 \bullet 7 - 2 \bullet 2 \bullet 1 \bullet 9 - 1 \bullet 3 \bullet 2 \bullet 11 - 1 \bullet 5 \bullet 4 \bullet 15] / 18}$$

or $\sigma_S = \sqrt{567} = 23.81$. Notice that if the data set could have been measured with sufficient precision (including a lower detection limit) so that no ties existed, then $\sigma_S = \sqrt{589.333} = 24.28$. Thus the ties here represent a rather small loss of information.

Example 2:

The example 1 data are modified to include ties, as follows:

y	1.22	2.20	4.80	1.28	1.97	1.97	2.64	2.34	4.84	2.96
x	2	24	99	99	377	544	632	3452	6587	53,170

Using a 0 to denote a tie, the comparisons used to compute P, M, and S are:

+	+	0_x	+	0_y	+	–	+	–
+	–	–	+	+	+	+	+	
+	–	–	+	+	+	+		
+	–	–	+	+	+			
+	+	–	+	+				0_x : tie in x
+	+	+	+					0_y : tie in y
+	+	–						
+	+							
+								

In total there are 33 +'s (P=33) and 10 –'s (M=10). Therefore S = 33–10 = 23, and $\tau = 23/45 = 0.51$. The exact two-sided p-value from table B8 is $2 \bullet 0.023 = 0.046$. For the large sample approximation, there are 2 ties of extent 2, so that

$$\sigma_S = \sqrt{[10 \bullet 9 \bullet 25 - 2 \bullet 2 \bullet 1 \bullet 9] / 18} = \sqrt{123} = 11.09$$

whereas without the tie σ_S was 11.18. Computing Zs,

$$Z_S = (23–1) / \sqrt{123}$$
$$= 22/(11.09) = 1.98.$$

From a table of the normal distribution, the 1-sided quantile for 1.98 = 0.976 so that $p \cong 2 \bullet (1–.976) = 0.048$.

8.3 Spearman's Rho

Spearman's rho is an alternative rank correlation coefficient to Kendall's tau. Kendall's tau is related to the sign test -- all positive differences between data pairs are assigned a +1 without regard to the magnitude of those differences. With Spearman's rho, differences between data values ranked further apart are given more weight, similar to the signed-rank test. Rho is perhaps easiest to understand as the linear correlation coefficient computed on the ranks of the data. Thus rho can be computed as a rank transform method. Rho and tau use different scales to measure the same correlation, much like Centigrade and Fahrenheit measures of temperature. Though tau is generally lower than rho in magnitude, their p-values for significance should be quite similar when computed on the same data.

To compute rho, the data for the two variables are ranked independently among themselves. For the ranks of x (Rx_i) and ranks of y (Ry_i), rho can be computed from the equation:

$$\text{rho} = \frac{\sum_{i=1}^{n} \left(Rx_i Ry_i\right) - n\left(\frac{n+1}{2}\right)^2}{n(n^2-1)/12} \qquad [8.5]$$

where (n+1)/2 is the mean rank of both x and y. Ties in x or y are assigned average ranks. This equation can be derived from substituting Rx_i and Ry_i for x_i and y_i in equation 8.6 for Pearson's r, and simplifying. If there is a positive correlation, the higher ranks of x will be paired with the higher ranks of y, and their product will be large. For a negative correlation the higher ranks of x will be related to lower ranks of y, and their product will be small. When there is no correlation, there will be nothing other than a random pattern in the association between x and y ranks, and their product will be similar to the product of their average rank, the second term in the numerator of equation 8.5. Thus rho will be close to zero.

Bhattacharyya and Johnson (1977) present the exact and large sample approximation versions of the hypothesis test for Spearman's rho. However, it is easiest to rank the two variables and compute the hypothesis test for Pearson's r -- the rank transform method. It is important to note that the large sample and rank approximations for rho do not fit the distribution of the test statistic well for small sample sizes (n<20), in contrast to Kendall's tau. This is one reason tau is often preferred over rho.

Example 1, continued

For the example 1 data, the data ranks are

Ry	1	5	9	2	4	3	7	6	10	8
Rx	1	2	3	4	5	6	7	8	9	10

Solving for rho, multiplying the ranks above gives,

$(Rx_i \bullet Ry_i)$ | 1 10 27 8 20 18 49 48 90 80, $\Sigma = 351$

$$\text{Rho} = \frac{351 - 10\,(5.5)^2}{1099/12} = \frac{48.5}{82.5}$$

= 0.588, exact p-value = 0.04 from table 13 of Bhattacharyya and Johnson (1977).

The approximate significance test for Pearson's r on the data ranks (as described in the next section) has a p-value = 0.074, not too close to the exact value. Whenever using Spearman's rho for sample sizes less than 20, exact p-values should be used.

8.4 Pearson's r

The most commonly-used measure of correlation is Pearson's r. It is also called the linear correlation coefficient because r measures the **linear** association between two variables. If the data lie exactly along a straight line with positive slope, then r = 1. This assumption of linearity makes inspection of a plot even more important for r than for rho or tau because a non-significant value of r may be due to curvature or outliers as well as to independence. As in figure 8.1, x and y may be strongly related in a nonlinear fashion, while the resulting r may be small and insignificantly different from zero.

Pearson's r is not as resistant to outliers as was tau and rho because it is computed using non-resistant measures -- means and standard deviations. It assumes that the data follow a bivariate normal distribution. With this distribution, not only do the individual variables x and y follow a normal distribution, but their joint variation also follows a specified pattern. This assumption rules out the use of r when the data have increasing variance, as in figure 8.1. Skewed variables often demonstrate outliers and increasing variance. Thus r is often not useful for describing the correlation between untransformed hydrologic variables.

Pearson's r is invariant to scale changes, as in converting streamflows in cubic feet per second into cubic meters per second, etc. This dimensionless property is obtained by standardizing, dividing the distance from the mean by the sample standard deviation, as shown in the formula for r, below.

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right) \qquad [8.6]$$

The significance of r can be tested by determining whether r differs from zero. The test statistic t_r is computed by equation 8.7, and compared to a table of the t distribution with n−2 degrees of freedom.

$$t_r = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \qquad [8.7]$$

Example 1, continued

For the example 1 data, means and standard deviations are:

	x	y
mean	6508.6	2.57
s	16531.6	1.31

Then $r = \frac{1}{9}\sum_{i=1}^{9}\left(\frac{x_i - 6508.6}{16531.6}\right)\left(\frac{y_i - 2.57}{1.31}\right) = 0.174$

To test for whether r is significantly different from zero, and therefore y is linearly dependent on x,

$$t_r = \frac{0.174\sqrt{8}}{\sqrt{1-(0.174)^2}} = 0.508,$$

with a p-value of 0.63 from a table of the t-distribution. Therefore H_0: r=0 is not rejected, and y is not linearly dependent (or related) to x as measured by r. This differs from the results using rho and tau, whose p-values of 0.04 and 0.07 respectively did indicate an association between y and x. Figure 8.4 provides an intuitive explanation of why r differs from rho and tau -- r is strongly affected by the one outlying observation, even though the overall trend is a linear one.

Exercises

8.1 Are uranium concentrations correlated with total dissolved solids in the following groundwater samples? If so, describe the strength of the relationship.

Uranium conc. in ppb	TDS, in mg/L	Uranium conc. in ppb	TDS, in mg/L
682.65	0.9315	1240.81	6.8559
819.12	1.9380	538.35	0.4806
303.76	0.2919	607.75	1.1452
1151.40	11.9042	705.89	6.0876
582.42	1.5674	1290.57	10.8823
1043.39	2.0623	526.09	0.1473
634.84	3.8858	784.68	2.6741
1087.25	0.9772	953.14	3.0918
1123.51	1.9354	1149.31	0.7592
688.09	0.4367	1074.22	3.7101
1174.54	10.1142	1116.59	7.2446
599.50	0.7551		

8.2 Compute the other two correlation coefficients not chosen in Exercise 8.1. Are all coefficients equally appropriate?

8.3 For the data on Corbicula densities in the Tennesse River found in Appendix C8, compute Kendall's tau for all pairs of data in the same strata and season, but one year apart. Is this correlation significant? How should this result be interpreted?